# Assignment -3
## Machine Learning.

Akshay Tambe (apt321@nyu.edu)
Snahil Singh (ss11381 @nyu.edu).

Q1. $$E(S) = -\sum_{l \in L} \frac{N_l}{N} \log_2 \frac{N_l}{N}$$

$$IG(S) = E(S) - \sum_{v \in V} \frac{|S_v|}{|S|} E(S_v)$$

a). $A \rightarrow 4+, 5-$

$B \rightarrow 3+, 6-$

$$E(S_A) = -\left\{ \frac{4}{9} \log_2 \frac{4}{9} + \frac{5}{9} \log_2 \frac{5}{9} \right\}$$

$$E(S_B) = -\left\{ \frac{3}{9} \log_2 \frac{3}{9} + \frac{6}{9} \log_2 \frac{6}{9} \right\}$$

Or the example with (4+, 5-) i.e A will have higher entropy than B because it's more balanced than B, or $N_+ / (N_+ + N_-)$ is closer to 1/2 for A than B. A has less disimilarity & less impurity than B.

b). $$IG(x_1) = E(S) - \sum_{v \in V} \frac{|S_v|}{|S|} E(S_v)$$

$N(F) = 4, \; N(T) = 3$

$$E(S) = -\left( \frac{3}{7} \log_2 \frac{3}{7} + \frac{4}{7} \log_2 \frac{4}{7} \right) = 0.52 + 0.46 = 0.98522$$

$$E(S_{x_1=T}) = -\left( \frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3} \right) = -(-0.52832 + (-0.3899)) = 0.918295830$$

$$E(S_{x_1=F}) = -\left( \frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4} \right) = -\log_2 \frac{2}{4} \quad 1$$

$$IG = 0.98522 - \left[ \frac{3}{7} \times 0.91829 + \frac{4}{7} \times 1 \right] = 0.020244.$$

c) . From question b,

Entropy $(S) = 0.985228139$

Entropy $(S_{x_1=T}) = -\left(\frac{1}{3}\log_2 \frac{1}{3} + \frac{2}{3}\log_2 \frac{2}{3}\right) = 0.918$

Entropy $(S_{x_1=F}) = -\left(\frac{3}{4}\log_2 \frac{3}{4} + \frac{1}{4}\log_2 \frac{1}{4}\right) = 0.8113$

Entropy $(S_{x_2=T}) = -\left(\frac{1}{3}\log_2 \frac{1}{3} + \frac{2}{3}\log_2 \frac{2}{3}\right) = 0.918$

Entropy $(S_{x_2=F}) = -\left(\frac{3}{4}\log_2 \frac{3}{4} + \frac{1}{4}\log_2 \frac{1}{4}\right) = 0.8113$

Entropy $(S_{x_3=T}) = -\left(\frac{1}{3}\log_2 \frac{1}{3} + \frac{2}{3}\log_2 \frac{2}{3}\right) = 0.918.$

Entropy $(S_{x_3=F}) = -\left(\frac{3}{4}\log_2 \frac{3}{4} + \frac{1}{4}\log_2 \frac{1}{4}\right) = 0.8113$

Information gain $(x_1) = 0.985228139 - \left[\frac{3}{7} \times 0.918 \cancel{0} + \frac{4}{7} \times 0.8113\right]$

$= 0.1281995$

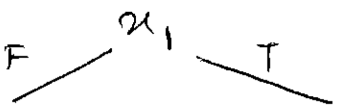
Similarly

Information gain $(x_2) = 0.1281995$ &

Information gain $(x_3) = 0.1281995$

Comparing all the info gain for $x_1, x_2, x_3$, we observe that all are equal

Hence, splitting on $x_1$



| | $x_1$ | $x_2$ | $x_3$ | $x$ |
|---|---|---|---|---|
| $x^1$ | F | F | F | + |
| $x^3$ | F | F | T | − |
| $x^4$ | F | T | F | + |
| $x^7$ | F | F | F | + |

| | $x_1$ | $x_2$ | $x_3$ | $x$ |
|---|---|---|---|---|
| $x^2$ | T | F | T | + |
| $x^5$ | T | T | F | − |
| $x^6$ | T | T | T | − |

Calculation of entropy & Info-gain for left subtree

Entropy $(S'_{x_2=T}) = -(\frac{1}{1}\log\frac{1}{1}+0)=0$

Entropy $(S_{x_2}'=F) = -(\frac{2}{3}\log\frac{2}{3}+\frac{1}{3}\log\frac{1}{3})$
$=0.918$

Info gain $(x_2) = -(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4})$
$= -(\frac{3}{4}\times 0+\frac{1}{4}\times 0.918)$

Info-gain $(x_2) = 0.5813$

Entropy $(S_{x_3}' = T) = -(0+\frac{1}{1}\log\frac{1}{1})=0$

Entropy $(S_{x_3}'=F) = -(\frac{3}{3}\log\frac{3}{3}+0)=0$

Info-gain $(x_3) = -(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4})-0$

∴ Infogain $(x_3) = 0.8113$

Here IG $(x_3) >$ IG $(x_2)$
hence splitting left subtree on $x_3$

Calculation of entropy & info-gain for right subtree

Entropy $(S_{x_2}'=T) = -(0+\frac{2}{2}\log_2\frac{2}{2})=0$

Entropy $(S_{x_2}'=F) = -(\frac{1}{1}\log\frac{1}{1}+0)=0$

Info-gain $(x_2) = -(\frac{1}{3}\log\frac{1}{3}+\frac{2}{3}\log\frac{2}{3})$
$-(\frac{1}{3}\times 0+\frac{2}{3}\times 0)$

Info-gain $(x_2) = 0.918$

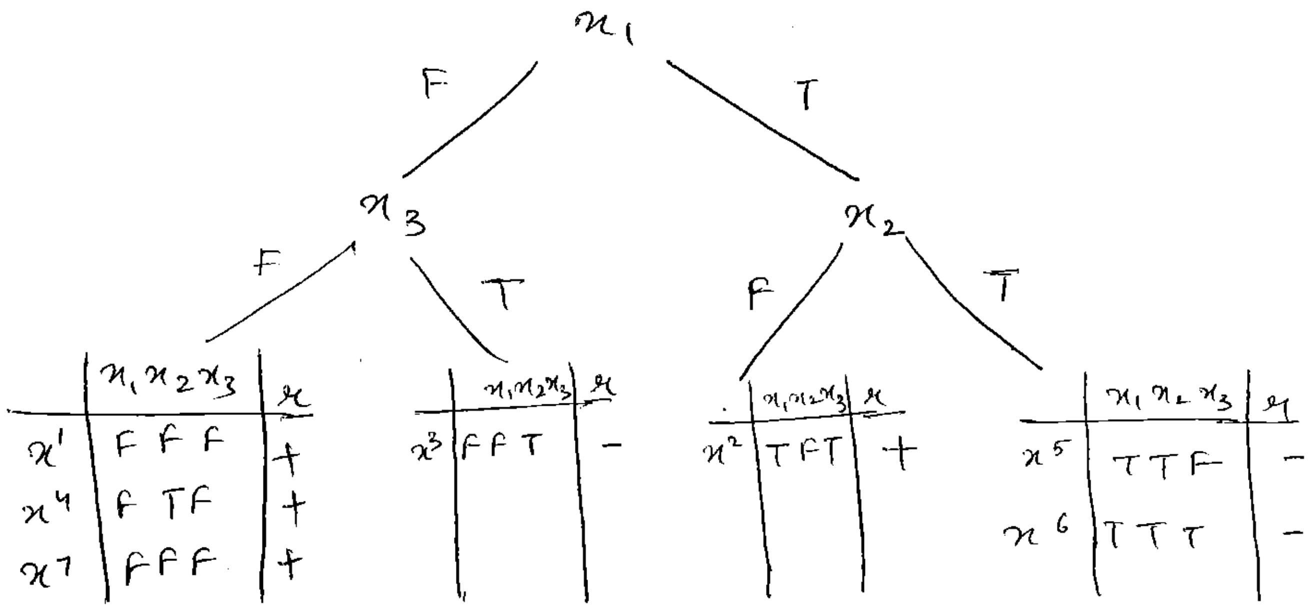
Entropy $(S_{x_3}' = T) = -(\frac{1}{2}\log\frac{1}{2}+\log\frac{1}{2})$
$=1$

Entropy $(S_{x_3}'=F) = -(0+\frac{1}{1}\log\frac{1}{1})=0$

Info-gain $(x_3) = 0.918 - [\frac{1}{3}\times 1+\frac{2}{3}\times 0]$

Info gain $(x_3) = 0.58467$

Here, IG $(x_2) >$ IG $(x_3)$ hence splitting right subtree on $x_2$



| | $x_1 x_2 x_3$ | $r$ |
|---|---|---|
| $x^1$ | F F F | + |
| $x^4$ | F T F | + |
| $x^7$ | F F F | + |

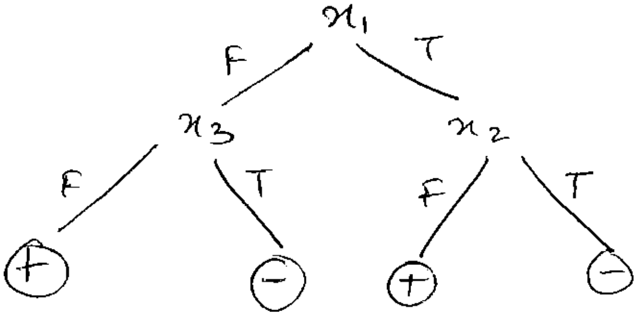
| | $x_1 x_2 x_3$ | $r$ |
|---|---|---|
| $x^3$ | F F T | - |

| | $x_1 x_2 x_3$ | $r$ |
|---|---|---|
| $x^2$ | T F T | + |

| | $x_1 x_2 x_3$ | $y$ |
|---|---|---|
| $x^5$ | T T F | - |
| $x^6$ | T T T | - |

As entropy is 0 by splitting on $x_3$ on left subtree & $x_2$ on right subtree, we can clearly see from observations

Final tree -



d) $\cancel{E}\,H(X) = -\sum_{i=1}^{n} P[X=i] \log_2 P[X=i] \rightarrow$ expected no. of bits needed to encode

$$H(Y|X) = \sum_{x} P[X=x] * \left( \sum_{Y} -P[Y=y|X=x] * \log_2 P[Y=y|X=x] \right)$$

M.I b/w X and Y = $H(\cancel{Y}|X)\cancel{=}\cancel{\Sigma}$ $H(Y) - H(Y|X) = H(X) - H(X|Y$

X → value of attribute $x_i$ $\{T, F\}$

Y → Label of example → +/−

Need to compute $H(X)$ and $H(Y|X)$, $H(X) - H(Y|X)$

$$H(Y|X) = P(T)\left[-P(+|T)\log_2 P(+|T) - P(-|T)\log_2 P(-|T)\right]$$
$$+ P(F)\left[-P(+|F)\log_2 P(\cancel{P}+|F) - P(-|F)\log_2 P(-|F)\right]$$

$$= \frac{3}{7}\cancel{\left[-\frac{3}{4}\log_2\right.} \quad \frac{3}{7}\cancel{\left[\frac{3}{3}\log_2\right.}$$

$$H(Y|X) = -\frac{3}{7}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) - \frac{4}{7}\left(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\right)$$

$$= -\frac{3}{7}(-0.92) = 0.96.$$

$$H(X) = -\left\{\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right\} = 0.98.$$

$$H(X) - H(Y|X) = 0.02.$$

~~d~~ e) Y → z possible labels.

Entropy $= -\sum \frac{N_l}{N} \log \frac{N_l}{N}$

Since all labels appear equally, we have $P(x) = 1/z$

$$= -\sum_{z} \frac{1}{z}\log\frac{1}{z} = -\log 1/z$$